# *Data*: Novartis GNF Liver Stage Dataset: Malariabox Annotation (AID: 602156)

# Number of active compounds: 274

# Number of inactive compounds: 5389

# *Input data*: the bioassay data is downloaded directly from pubchem database using a short R script. By using different rcdk functions, information about 1)smiles 2)CID 3)activity are extracted from input data and are put together as a table in SMILES format. (R script: “getactivity.r”| input file: “602156want.smi” )

# *Predictive modeling*: (R script: modeling.r)

# 3 models: Random Forest, Naïve Bayes, and support vector machine (SVM)

# 4 fingerprints: maccs (166), pubchem (88), extended (depth=6, size=1024), estate (79)

# model validation: k-fold validation (in this case: 5-fold )

# analysis of model performance: ROC curve, area under curve (CUV, ranging from 0.5 to 1.0, the larger the better), sensitivity (true positive rate, tpr), specificity (true negative rate, tnr), accuracy, precision, F1 score.

# Fingerprint: maccs (166)

# 

# Fingerprint: pubchem (881)

# 

# Fingerprint: extended (depth=6, size=1024)

# 

# Fingerprint: estate (79)

# 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC value | maccs | pubchem | extended | estate |
| Random Forest | 0.72\* | 0.85\* | 0.76\* | 0.91 |
| Naïve Bayes | 0.72\* | 0.50 | 0.50 | 0.97\* |
| SVM | 0.65 | 0.74 | 0.57 | 0.73 |

# \* indicates the highest score of the model in the particular type of fingerprint

# # indicates the highest score of the fingerprint in the same model

# As we can see through the table, for all the 4 types of fingerprints, the predictive model generated by Random Forest always has the highest AUC value compared to Naïve Bayes and SVM. On the other side, with the same model, estate fingerprints always has the highest AUC value compared to the other three fingerprints. (There is an exception in pubchem fingerprint where its AUC value is 0.74 which is only 1% higher than 0.73 of estate AUC score.)

# Estate fingerprint is the descriptor. This descriptor corresponds to the E-State atom types described by Hall and Kier. As a result, topological indices show information about the nature of chemical structure graph which can explains why estate fingerprints may be a better choice when building predictive model: it provides more information about the structure of compounds.

# In Random Forests, when training the data, a multitude of decision trees are constructed and and the output is the model generated by the individual tree. As SVM uses two possible input classes to output model, it is more useful in problems like classification or grouping. Before using Naïve Bayes, we have to remember that there is a strong assumption that the data are independent. In our case, as an ensemble learning method, Random Forests seems better suit to build the compound predictive model.